

Comment

William Navidi

I would like to address some of the issues raised by Professor Young, then add an example to the interesting collection presented in his paper.

Professor Young draws a distinction between the i.i.d. case, where bootstrap methods are often simple and automatic, and the nonindependent case, where, if the bootstrap can be made to work at all, the appropriate method is likely to be rather sophisticated and specific to the problem at hand. It seems to me that this situation may be unavoidable, in that the appropriate choice of bootstrap method is often determined by the dependence structure of the data, with the i.i.d. case being a particularly simple form of dependence structure yielding a particularly simple bootstrap method.

For example, let X_1, \dots, X_n be i.i.d. Bernoulli trials with success probability p . Let $\hat{p} = (1/n)\sum_{i=1}^n X_i$ be the usual estimator of p . We can use the bootstrap to estimate the distribution of \hat{p} by defining X_1^*, \dots, X_n^* to be i.i.d. Bernoulli trials with success probability \hat{p} . This is the usual bootstrap method for i.i.d. data, since the random variables X_1^*, \dots, X_n^* are an i.i.d. sample from the empirical distribution of X_1, \dots, X_n . If we define $\hat{p}^* = (1/n)\sum_{i=1}^n X_i^*$, then the asymptotic distribution of $(\hat{p}^* - \hat{p})$ is the same as that of $(\hat{p} - p)$. Now assume instead that X_1, \dots, X_n are Bernoulli trials with common success probability p , generated from a Markov process with transition matrix P . If we define X_1^*, \dots, X_n^* as above, the bootstrap will fail, because the variance of \hat{p} depends on P . In order for the bootstrap to be successful, the bootstrap distribution must be based on an estimate \hat{P} of P , not just on the marginal estimate \hat{p} . This shows that in general the bootstrap must be tailored to the dependence structure of the data in order to produce good results. It seems that the degree to which universal methods can be developed may be limited.

I am glad that Professor Young has pointed out examples where asymptotic accuracy does not carry over to small samples. In particular, the role of higher-order asymptotics in the study of the boot-

strap needs to be assessed. The bootstrap is often accurate to a higher order than conventional approximations based on the central limit theorem, essentially because the bootstrap consistently estimates the third and higher moments of a distribution while the normal approximation does not. This has resulted in much beautiful theory. What needs more investigation is the extent to which higher-order asymptotic accuracy translates into better accuracy in small samples. This of course is what is important in practice, not only because real samples are finite, but because one is inevitably faced with errors of size $O(1)$ due to sampling bias and model misspecification, so reducing the remaining error from $O(1/\sqrt{n})$ to $O(1/n)$ may not matter very much.

A topic mentioned only briefly in the paper which is potentially quite important is the use of the bootstrap in model selection or in assessing the accuracy of model selection procedures. Two of the papers in LePage and Billard (1992), by Kipnis (1992) and by Brownstone (1992), address this issue. Discussions of this topic can also be found in Efron (1983), Gong (1986) and Freedman, Navidi and Peters (1988). To give a concrete example, we will discuss variable selection in a linear regression model. The model is $Y = X\beta + \varepsilon$, where the matrix X has n rows and p columns, and ε is a vector of i.i.d. errors with mean 0 and variance σ^2 . If it is thought that the matrix X has a large number of columns which are not well related to the dependent variable Y , a selection procedure may be used in an attempt to create a model of manageable size that retains most of the explanatory power of the original. Variable selection is a widely practiced art, with procedures such as forward selection, stepwise selection, all subsets, Mallows C_p and others being found in many statistical software packages.

It is well known that standard goodness-of-fit statistics applied to selected models often seriously exaggerate the degree of fit, because the selected variables tend to be those whose sample correlation with Y is inflated by random error. It is worthwhile to investigate whether the bootstrap can be used to obtain more realistic estimates of fit. Assume for technical reasons that the matrix X contains an intercept. The standard method of applying the bootstrap in ordinary least squares regression is as follows (Efron,

William Navidi is Associate Professor of Biostatistics, Department of Preventive Medicine, University of Southern California, Los Angeles, California 90033.

1979a; Freedman, 1981):

1. Compute $\hat{\beta} = (X^T X)^{-1} X^T Y$.
2. Let e_1, \dots, e_n be the residuals $e = Y - X\hat{\beta}$.
3. Let e_1^*, \dots, e_n^* be an i.i.d. sample from the empirical distribution of e_1, \dots, e_n .
4. The bootstrap model is $Y^* = X\hat{\beta} + e^*$.

The bootstrap model is much like the real model, with the advantage that the “true” value of β , namely, $\hat{\beta}$, is known. The bootstrap model works for inference about the distribution of $\hat{\beta}$ in that if $\hat{\beta}^* = (X^T X)^{-1} X^T Y^*$, then, under mild conditions on the rate of growth of the elements of X , the asymptotic distribution of $(\hat{\beta}^* - \hat{\beta})$ is the same as that of $(\hat{\beta} - \beta)$ [see Freedman (1981)]. It might be hoped that this would enable the bootstrap model to reflect accurately the behavior of estimates based on selected columns of X as well. Unfortunately, this does not seem to be the case. Roughly speaking, this is be-

cause $\mathbb{E}\|X\hat{\beta}\|^2 > \|X\beta\|^2$, while $\text{Var}(e_i^*) = (1/n)\sum_{i=1}^n \mathbb{E}(e_i^2) = (n-p)\sigma^2/n$. In other words, the mean of Y^* tends to be larger than that of Y , while its variance tends to be less. Thus the bootstrap model tends to confirm the overoptimistic assessment of goodness of fit produced by model selection. The asymptotic performance of the bootstrap is good as $n \rightarrow \infty$ with p fixed, since $(1/n)\|X\hat{\beta}\|^2 \rightarrow (1/n)\|X\beta\|^2$ under mild conditions on the rate of growth of the elements of X . When p is a substantial fraction of n however, which is often the case in variable selection, the results can be quite misleading (Freedman, Navidi and Peters, 1988). Potential solutions may involve shrinking the length of $\hat{\beta}$ for use in the bootstrap model. Since the use of model selection procedures is quite extensive in statistical practice, better methods of assessing the performance of selected models would be very useful. I think it is likely that the bootstrap will turn out to have something to offer in this area.

Comment

Mark J. Schervish

Professor Young is to be congratulated on summarizing so succinctly and clearly the vast body of work on the bootstrap which has appeared since 1979. The bootstrap has achieved a remarkable level of notoriety both due to its analytical simplicity and to its seeming ability to serve up the proverbial “free lunch.” However, behind all of the technical details of the bootstrap and its asymptotics, there still lies the question of why does (or does not) the bootstrap work in general. The theoretical use of the bootstrap involves the replacement of a distribution F in a formula $T(X, F)$ by some other distribution \hat{F} . The degree to which this replacement is successful depends on the degree to which \hat{F} resembles F in important regards. For example, suppose that F is a distribution with finite variance, \hat{F} is the empirical distribution and $T(X, F)$ is the average \bar{X} of the sample X minus the mean of the distribution F . Then the variance of $T(Y, \hat{F})$ (where Y is a sample from \hat{F}) can be expected to be a lot like the variance of $T(X, F)$. On the other hand, if F is a continuous distribution on an interval $[0, \theta]$ and $T(X, F) = n(\theta - X_{(n)})$, where

$X_{(n)}$ is the largest order statistic, then Young points out the well-known fact that $\Pr(T(Y, \hat{F}) = 0)$ converges to $1 - \exp(-1)$ as $n \rightarrow \infty$, while $T(X, F)$ has a continuous distribution.

I believe that some insight into what the bootstrap does can be gained by doing something with this last example that is uncommon in most bootstrap applications, namely, that we think about the problem. An obvious observation is that \hat{F} and F differ markedly in the manner in which the largest order statistic from a sample is related to the least upper bound on the support of the distribution. In particular, with \hat{F} , the two can be equal with non-negligible probability; with F , they cannot. An obvious, albeit naive, response is to smooth \hat{F} , that is, replace the empirical distribution by a continuous distribution which approximates it. For example, if $X_{(1)} \leq \dots \leq X_{(n)}$ are the order statistics, one could define $\hat{F}(x) = G(u)i/n + [1 - G(u)](i-1)/n$ for $X_{(i-1)} < x < X_{(i)}$, where G is a continuous distribution function and $u = (x - X_{(i-1)})/(X_{(i)} - X_{(i-1)})$. (Forget about $x < X_{(1)}$ for now.) Bickel and Freedman (1981) claim that even this does not mend the problem. They attribute (page 1210) the problem to “the lack of uniformity in the convergence of” \hat{F} to F . In fact, it is not difficult to see what happens in this case. We get that $T(Y, \hat{F})$ is the sum of two random vari-

Mark J. Schervish is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.